DOCUMENT RESUME

ED 384 646 TM 023 859

AUTHOR Kirisci, Levent; Hsu, Tse-Chi
TITLE The Robustness of BILOG to Violations of the Assumptions of Unidimensionality of Test Items and Normality of Ability Distribution.
PUB DATE Apr 95
NOTE 29p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 19-21, 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Ability; Comparative Analysis; Correlation; Difficulty Level; *Estimation (Mathematics); Guessing (Tests); *Robustness (Statistics); Simulation; *Statistical Distributions; *Test Items
IDENTIFIERS ASCAL Computer Program; *BILOG Computer Program; *Dimensionality (Tests); MULTILOG Computer Program; Three Parameter Model; Unidimensionality (Tests)

ABSTRACT

The main goal of this study was to assess how sensitive unidimensional parameter estimates derived from BILOG were when the unidimensionality assumption was violated and the underlying ability distribution was not multivariate normal. A multidimensional three-parameter logistic distribution that was a straightforward generalization of the unidimensional three-parameter logistic distribution was used as a data generation model for simulations. The study found that test dimensionality had a significant effect on the estimation of item parameters and ability. The BILOG program was not robust against the violation of the underlying ability distribution in estimating the item guessing parameter and ability for a unidimensional or multidimensional test. If dimensions of ability were correlated, and a test was multidimensional, BILOG yielded more robust estimates of item difficulty, guessing, and ability. The shape of the ability distribution affected only the estimation of ability. Comparison with the ASCAL and MULTILOG programs suggested that no program performed better than the others in estimating item and ability parameters across all test dimensions, degrees of intercorrelations of ability dimensions, and all ability distributions. Seven tables and seven figures illustrate the discussion. (Contains 10 reference.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 384 646

THE ROBUSTNESS OF BILOG TO VIOLATIONS OF THE ASSUMPTIONS OF UNIDIMENSIONALITY OF TEST ITEMS AND NORMALITY OF ABILITY DISTRIBUTION

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
LEVENT KIRISCI
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Levent Kirisci, Ph.D.

and

Tse-Chi Hsu, Ph.D.

University of Pittsburgh

Pittsburgh Adolescent Alcohol Research Center

Department of Psychiatry

University of Pittsburgh School of Medicine

Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, April 1995

TM023859

BEST COPY AVAILABLE

Abstract

The main goal of this study was to assess how sensitive unidimensional parameter estimates derived from BILOG when unidimensionality assumption was violated and underlying ability distribution was not multivariate normal. More specifically, the study was designed to answer the following questions: (1) did the dimensionality of a test affect parameter estimation?, (2) did the shape of underlying ability distribution affect parameter estimation?, (3) did the degree of intercorrelation between ability dimensions affect parameter estimation?, and (4) finally, were there any significant interactions between two and more factors mentioned above? In addition, the results obtained from this study were compared with the results obtained from previous study which compared ASCAL and MULTILOG under the same conditions.

One crucial assumption in item response theory (IRT) is related to unidimensionality of test items, because (1) interpretations of a test score are more articulate, and (2) IRT has been developed under the assumption of unidimensionality, i.e., IRT assumes that items measure a single trait. In education and psychology, however, tests developed are seldomly unidimensional. As a result, applications of IRT are limited to tests that are predominantly unidimensional. There are two procedures to overcome this predicament: (1) to sort multidimensional items into unidimensional subtests and apply IRT to each subtest (Reckase & McKinley, 1985) or (2) to use a multidimensional IRT model that relates dichotomous item responses to the magnitude of an examinee's ability on each trait. Since its complexity, most multidimensional IRT models have not been developed. Therefore, applying univariate IRT models to multidimensional problems is a common practice. In this procedure, the accuracy of the item parameter and ability estimates are two major concern in computer programs. There are three widely used programs available for estimation of item and ability parameters such as, ASCAL (Assessment System Corporation, 1988; Vale & Gialluca, 1985), BILOG (Mislevy & Bock, 1990), and MULTILOG (Thissen, 1991). The robustness of violating the unidimensionality assumption on the quality of parameter estimation has been investigated by Yu and Hsu (1990). They found the existence of robustness of unidimensional IRT models depended on the intercorrelation between underlying abilities and also depended on the type of calibration program used, although violations of the unidimensionality assumption in most cases seriously affected parameter estimation. They also concluded that unidimensional IRT estimation procedures were generally not very sensitive to different shapes of underlying ability distribution. Thus, the main goal of this study was to assess how sensitive unidimensional parameter estimates derived from BILOG if unidimensionality assumption was violated and underlying ability distribution was not multivariate normal. More specifically, the study was designed to answer the following questions: (1) did the dimensionality of the test (unidimensional versus multidimensional) affect parameter estimation?, (2) did the shape of underlying ability distribution (normal versus skewed versus peaked) affect parameter estimation?, (3) did the degree of intercorrelation between ability dimensions (correlated versus uncorrelated) affect parameter estimation?, and (4) finally, were there any significant interactions between two and more factors

mentioned above?. In addition, results obtained from ASCAL were compared with MULTILOG and BILOG.

METHOD

In this study, a multidimensional three-parameter logistic distribution that was a straightforward generalization of the unidimensional three-parameter logistic distribution was used as a data generation model. This model was chosen because inferences from three dimensional to lower dimensional models and also to models with lower parameters could be easily done. The multidimensional three-parameter distribution can be written as

$$P_i(\underline{\theta}_i)=c_i + (1-c_i)/(1+\exp(-1.7\underline{a}_i'(\underline{\theta}_i-\underline{b}_i)))$$

where $P_i(\underline{\theta}_i)$ is the probability of a correct response to item j by examinee i, $\underline{\theta}_i = (\theta_{i1},\theta_{i2},...,\theta_{ik})$is the ability parameter vector for examinee i, k is the number of ability dimension,$\underline{a}_j = (a_{j1}, a_{j2}, ..., a_{jk})$ is the discrimination parameter vector for item j, $\underline{b}_j = (b_{j1}, b_{j2}, ..., b_{jk})$ is the difficulty parameter vector for item j, and c_j is the guessing parameter for item j (Hattie, 1991; Doody-Bogan and Yen, 1983).

The factors manipulated were: (1) test dimensionality; one- or three-dimensional test, (2) shape of underlying ability distribution; normal, positively skewed, or platykurtic, and (3) degree of intercorrelation between underlying abilities; 0 or 0.6. The root mean squared error between the true and estimated parameters was used as an outcome variable.

In this study, discrimination parameter was distributed uniformly on the interval of 0.4 and 2.0; difficulty parameter was assumed to be distributed as uniformly on the interval of -2.0 and 2.0; and guessing parameter was uniformly distributed on the interval of 0.0 and 0.3. The IMSL (1989) subroutines were used to generate the random numbers.

Three univariate ability distributions with zero mean and unit variance were specified as follows: (1) normal distribution (skewness=0.0 and kurtosis=0.0), (2) positively skewed distribution (skewness=0.75 and kurtosis=0.0), and (3) platykurtic distribution (skewness=0.0 and kurtosis= -1.0). The IMSL subroutines were used to generate normal deviates. Fleishman's (1978) power procedure was used to generate nonnormal deviates: $Y=a+bz+cz^2+dz^3$ where Y was a nonnormal deviate with specified skewness and

kurtosis and z was normal deviate. The coefficients a, b, c, and d were known constants. To generate three-dimensional normal or nonnormal ability distribution with specified intercorrelations and skewness and kurtosis, X=AY transformation was used where $Y'=(Y_1,Y_2,Y_3)$, $V=AA'$ and A was the Cholesky factorization of the desired covariance matrix V of X,

$$\text{V is either } \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \text{ or } \begin{bmatrix} 1 & .6 & .6 \\ .6 & 1 & .6 \\ .6 & .6 & 1 \end{bmatrix}, \text{ then } A=\begin{bmatrix} 1 & 0 & 0 \\ .6 & .8 & 0 \\ .6 & .3 & .7416 \end{bmatrix}.$$

To generate three trivariate ability distribution with intercorrelation 0.6 and variance of 1 on each dimension, the random vector X_j was simulated where

$$X_j = AY = \begin{bmatrix} 1 & 0 & 0 \\ .6 & .8 & 0 \\ .6 & .3 & .7416 \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix},$$

where Y_1, Y_2, Y_3 were independent univariate random variables with either normal(0,1), positively skewed or platykurtic distributions.

To generate item responses, the probability of a correct response of each examinee to each item under each experimental condition was computed and then compared with a uniform random number between 0 and 1 to create a (0,1) item response matrix of order 1000 x 40, where 40 was the number of items. Each experimental condition was replicated 10 times. In this simulation study, BILOG was used to calibrate items and estimate ability levels.

Adequacy of generated data. Table 1 and Table 2 summarize the distribution of ability parameters. Mean, standard deviation, skewness, kurtosis, and correlations between abilities were averaged over 10 replications. Normal, positively skewed, and platykurtic data were generated for a total of 9 experimental combinations. Ten random samples for each combination were generated from populations possessing the desired distributional characteristics. As can be seen, the average scores were reasonably close to the theoretical values. In summary, data generation procedure was adequate.

Table 1 and Table 2

Table 3 describes the psychometric characteristics of the generated 40-item test. Items appeared to be in the middle range of difficulty (Range: 0.542 - 0.558). Reliability scores ranged between 0.923 and 0.959. Discriminative power varied between moderate to high.

Table 3

Design and data analysis. Three-way ANOVA was used to test the effects of test dimensionality (1-, 3-dimensional), the shape of ability distribution (normal, positively skewed, platykurtic), and their interaction. Two separate analyses were conducted: (1) the correlation between ability dimensions were 0 and (2) the correlation between ability dimensions were 0.6. The dependent variable was a root mean squared error (RMSE) term that was based on the difference between the average of true parameters and estimate. Later, three-way ANOVA was used second time for only 3-dimensional data to test the effects of degree of intercorrelation between ability dimensions, the shape of ability distribution, and their interaction. RMSE that was based on the difference between the average of true parameters and estimate was used as a dependent variable.

RESULTS

Table 4 and Table 5 present the mean and standard deviation of RMSE between the parameter estimates and true parameter values (true multidimensional parameter values or the average of the true parameters) for each experimental combination. As can be seen, RMSEs of discrimination and difficulty

were lowest for the 1-dimensional data across all ability distributions and degree of intercorrelations among dimensions of ability. RMSE of guessing for the 1-dimensional data was higher than the 3-dimensional correlated and uncorrelated data. However, RMSE of ability of 1-dimensional data was lower than the 3-dimensional uncorrelated data, and higher than the 3-dimensional correlated data when RMSE of average of true abilities was used as a dependent variable.

Table 4 and Table 5

Results of Three-way ANOVAs.

Table 6 summarizes the results of 3-way ANOVA for testing the effects of test dimensionality, types of ability distribution, and their interaction when intercorrelation between dimensions of ability was either 0 or 0.6. The dependent variable was RMSE that was calculated from the average of item parameters and estimates obtained from BILOG. Table 7 also summarizes the results of 3-way ANOVA for testing the effects of intercorrelations between dimensions of ability, ability distribution, and their interaction when test dimension was set to 3. Tests results were based on RMSE that was calculated as a difference between the average of true parameters and estimate.

Table 6 and Table 7

Item discrimination. As can be seen in Table 6, the effect of test dimension was significant regardless of degree of intercorrelation between dimensions of ability. The RMSE of 1-dimensional test was significantly smaller than RMSE of 3-dimensional test at 0.05 level. The shape of ability distribution and interaction of test dimension and intercorrelation between abilities did not have any significant effects.

Table 7 summarizes the results of 3-way ANOVA when test dimension was set to 3. If dimensions of ability were uncorrelated, RMSE of average of discrimination parameters was significantly smaller than

dimensions of ability were correlated.

Item difficulty. Test dimension was significant for both uncorrelated and correlated abilities. RMSE of 1-dimensional test was significantly smaller than RMSE of 3-dimensional test at 0.05 level.

In Table 7, the correlation effect was significant. When dimensions of ability were correlated, RMSE of correlated data was significantly smaller than the uncorrelated data at 0.05 level for a multidimensional test.

Item guessing. Test dimensionality and ability distribution had significant effects when correlation between underlying ability dimension was either 0 or 0.6. The results showed that the 3-dimensional test had significantly smaller RMSE than the 1-dimensional test at 0.05 level. Furthermore, the normal distribution produced significantly smaller RMSE than positively skewed distribution in both cases of uncorrelated and correlated abilities at 0.05 level. In Table 7, correlated abilities gave significantly smaller RMSE than the uncorrelated abilities at 0.05 for a 3-dimensional test.

Ability estimates. In Table 6, interaction effects were observed between test dimension and ability distribution when underlying abilities were uncorrelated and correlated. If underlying correlation between abilities was 0, 1-dimensional test produced significantly smaller RMSE across all ability distributions as illustrated in Figure 1. If underlying correlation was 0.6, 3-dimensional test produced smaller RMSE across ability distributions than 1-dimensional test. In Table 7, correlation among dimensions of ability and the ability distribution had significant interaction effect. If correlation between abilities was 0.6, then the resulting RMSE was significantly smaller regardless of ability distribution, as shown in Figure 2.

Figure 1 and Figure 2

Comparison of BILOG with ASCAL and MULTILOG.

In the final section, the results of this study were compared with the results obtained from ASCAL and MULTILOG (Yu & Hsu, 1990). RMSEs of item parameters and ability were plotted with respect to BILOG, ASCAL, and MULTILOG. Figure 3 shows that BILOG produced smaller RMSEs of item parameters regardless of test dimension (1- or 3-dimensional). However, the RMSE of ability obtained from

BILOG was smaller for only 1-dimensional test. If the correlation between ability dimensions was 0.60, BILOG yielded smaller RMSEs for all item parameters and ability for both 1- and 3-dimensional tests as shown in Figure 4, except in estimating item discrimination parameter for a 3-dimensional test.

Figure 3 and Figure 4

Figure 5 and Figure 6 illustrate the performance of 3 computer programs with respect to ability distributions. When correlation between ability dimensions was 0, BILOG did not perform better for estimating item discrimination, item difficulty, and ability. However, the RMSE of item guessing was smaller across all ability distributions. When the correlation between ability dimensions was 0.6, BILOG estimated item discrimination parameter worse than ASCAL and MULTILOG. In estimating difficulty, guessing, and ability, BILOG performed better or produced similar results.

Figure 5 and Figure 6

In Figure 7, three programs were compared with respect to RMSEs they produced when dimensions of ability were uncorrelated and correlated. If dimensions of ability were uncorrelated, BILOG yielded smaller or comparable RMSEs for item parameters as in Figure 7 with the exception of item discrimination parameter when abilities were correlated.

Figure 7

SUMMARY AND CONCLUSION.

This study attempted to clarify the effects of applying unidimensional IRT models to calibrate three-

dimensional data generated from a multidimensional model with different underlying ability distributions. The results of the study provided valuable information regarding the applicability and feasibility of using unidimensional methods in a multidimensional model. It determined (1) the effects of dimensionality on item and ability parameter estimation, (2) the effects of the shape of underlying ability distribution, (3) the effects of intercorrelations between ability dimensions, and (4) finally, any interaction effect of any two or more factors on parameter and ability estimation.

Results related to robustness of BILOG to vilolations of the assumptions of unidimensionality of test items and normality of ability distribution were summarized as follows: (1) This study found that test dimensionality had a significant effect on estimation of item parameters and ability. One-dimensional test performed better in estimating item discrimination and difficulty. In contrast, 3-dimensional test produced better estimates of guessing parameter regardless of intercorrelation among abilities. It is noteworty that an unidimensional test produced better estimates of item discrimination, difficulty, and ability, if dimensions of ability were uncorrelated. This indicated that the assumption of unidimensionality of IRT model was crucial in item calibration and ability estimation; (2) BILOG was not robust against the violation of underlying ability distribution in estimating item guessing parameter and ability for either unidimensional test or multidimensional test. Positively skewed ability distribution gave the poorest estimates of ability and item guessing parameter when abilities correlated or uncorrelated; (3) If dimensions of ability were correlated and a test was multidimensional, BILOG yielded more robust estimates of item difficulty, guessing and ability; (4) Although there were significant interaction effects, they did not provide additional information. In summary, violations of test dimensionality and degree of intercorrelation among dimensions of ability had crucial effects on the estimation of item parameters and ability. Furthermore, the shape of the ability distribution affected only the estimation of ability.

The comparison of BILOG with ASCAL and MULTILOG were summarized as follows: (1) BILOG was at least as robust as ASCAL and MULTILOG in estimating item parameters for 1-dimensional and 3-dimensional tests regardless whether ability dimensions were correlated or not; (2) If dimensions of ability were uncorrelated, MULTILOG performed better in estimating item discrimination parameter when underlying ability distribution was normal and platykurtic; three programs performed similarly when the

distribution was positively skewed. ASCAL accomplished better results in estimating item difficulty and ability parameters across 3 ability distributions. However, BILOG was more robust in estimating item guessing parameter than other programs for all ability distributions. If dimensions of ability were correlated, BILOG gave more robust estimates for item guessing and ability parameters across 3 ability distributions. BILOG performed poorly for estimating item discrimination parameter for all ability distributions. In estimating item difficulty parameter, BILOG produced more robust estimates for positively skewed and platykurtic ability distribution; (3) Finally, BILOG's performance was better or comparable in estimating item discrimination, item difficulty, item guessing parameters and ability, If dimensions of ability were uncorrelated. BILOG yielded more robust estimates in estimating item difficulty, item guessing, and ability parameters when ability dimensions were correlated.

In short, these results suggested that no program performed better than the others in estimating item and ability parameters across all test dimensions, degree of intercorrelations among ability dimensions, and all ability distributions. Thus, user should be very attentive in choosing a program to calibrate items or estimating ability parameter and should take into a consideration the shape of the ability distribution, the degree of intercorrelation among abilities, and test dimensionality.

REFERENCES

Assessment System Corporation. (1988). User's manual for the MicroCAT testing System. St.Paul, MN: Author.

Doody-Bogan, E., & Yen, W.M. (1983). Detecting multidimensional and examining its effects on vertical equating with the three-parameter logistic model. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Fleishman, A.I. (1987). A method for simulating nonnormal distributions. Psychometrika, 43, 521-532.

Hattie, J.A. (1981). Decision criteria for determining unidimensionality. Unpublished doctoral dissertation, University of Toronto.

IMSL Library (10th edition). (1989). International Mathematical and statistical libraries. Houston, TX: Author.

Mislevy, R.J., & Bock, R.D. (1990). BILOG: Maximum likelihood item analysis and test scoring with logistic models for binary items. Mooresville, IN: Scientific Software, Inc.

Reckase, M.D., & McKinley, R.L. (1985). Some latent trait theory in a multidimensional latent space. In D.J. Weiss (Ed.), Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology.

Thissen, D. (1991). MULTILOG Version 6.0 User's Guide. Mooresville, IN: Scientific Software, Inc.

Vale, C.D., & Gialluca, K.A. (1985). ASCAL: A microcomputer program for estimating logistic IRT item parameters (Research report No. ONR-85-4). St. Paul, MN: Assessment Systems Corporation.

Yu, L., & Hsu, T.C. (1990). The robustness of ASCAL and MULTILOG to violations of the assumption of unidimensionality of test items and normality distribution. Paper presented at the annual meeting of AERA, Boston, April.

Table 1. Average mean (Mean), standard deviation (SD), skewness (SK), and kurtosis (KR) based on 10 replications and 1000 simulated examinees for given ability distribution and intercorrelation.

Ability Distribution	Statistic	Expected	Observed Mean (SD)
Normal (1-dimensional)	Mean	0.000	-0.021 (0.033)
	SD	1.000	0.982 (0.015)
	SK	0.000	0.005 (0.050)
	KR	0.000	0.035 (0.176)
Positively Skewed (1-dimensional)	Mean	0.000	-0.003 (0.021)
	SD	1.000	1.000 (0.028)
	SK	0.750	0.736 (0.054)
	KR	0.000	-0.065 (0.144)
Platykurtic (1-dimensional)	Mean	0.000	-0.028 (0.032)
	SD	1.000	1.001 (0.015)
	SK	0.000	0.044 (0.048)
	KR	-1.000	-0.986 (0.037)
Normal (3-dimensional; corr=0.0)	Mean	0.000	0.012 (0.029)
	SD	1.000	0.995 (0.020)
	SK	0.000	-0.016 (0.075)
	KR	0.000	-0.018 (0.101)
	Mean	0.000	0.013 (0.038)
	SD	1.000	1.008 (0.027)
	SK	0.000	0.010 (0.066)
	KR	0.000	0.077 (0.085)
	Mean	0.000	-0.009 (0.030)
	SD	1.000	1.002 (0.014)
	SK	0.000	0.031 (0.070)
	KR	0.000	0.093 (0.170)
Positively Skewed (3-dimensional; corr=0.00)	Mean	0.000	-0.011 (0.026)
	SD	1.000	0.991 (0.022)
	SK	0.750	0.762 (0.065)
	KR	0.000	0.027 (0.179)
	Mean	0.000	0.008 (0.018)
	SD	1.000	0.999 (0.018)
	SK	0.750	0.750 (0.059)
	KR	0.000	0.005 (0.100)
	Mean	0.000	-0.015 (0.025)
	SD	1.000	0.997 (0.017)
	SK	0.750	0.767 (0.039)
	KR	0.000	0.032 (0.101)
Platykurtic (3-dimensional; corr=0.00)	Mean	0.000	-0.003 (0.031)
	SD	1.000	1.001 (0.011)
	SK	0.000	-0.005 (0.039)
	KR	-1.000	-1.005 (0.035)
	Mean	0.000	0.000 (0.025)
	SD	1.000	1.005 (0.008)
	SK	0.000	0.000 (0.045)
	KR	-1.000	-1.014 (0.011)
	Mean	0.000	-0.008 (0.028)
	SD	1.000	0.996 (0.124)
	SK	0.000	0.006 (0.035)
	KR	-1.000	-0.996 (0.029)

BEST COPY AVAILABLE

Table 1 (continued)

Ability Distribution	Statistic	Expected	Observed Mean (SD)
Normal (3-dimensional; corr=0.60)	Mean	0.000	-0.000 (0.027)
	SD	1.000	0.990 (0.025)
	SK	0.000	-0.007 (0.061)
	KR	0.000	0.051 (0.177)
	Mean	0.000	0.005 (0.026)
	SD	1.000	1.000 (0.028)
	SK	0.000	-0.054 (0.050)
	KR	0.000	0.071 (0.209)
	Mean	0.000	0.006 (0.025)
	SD	1.000	1.002 (0.023)
	SK	0.000	-0.027 (0.063)
	KR	0.000	0.035 (0.113)
Positively Skewed (3-dimensional; corr=0.60)	Mean	0.000	0.004 (0.025)
	SD	1.000	0.994 (0.019)
	SK	0.750	0.741 (0.049)
	KR	0.000	-0.038 (0.102)
	Mean	0.000	0.000 (0.029)
	SD	1.000	0.998 (0.019)
	SK	0.750	0.728 (0.062)
	KR	0.000	0.034 (0.176)
	Mean	0.000	-0.001 (0.027)
	SD	1.000	0.991 (0.027)
	SK	0.750	0.748 (0.068)
	KR	0.000	0.031 (0.176)
Platykurtic (3-dimensional; corr=0.60)	Mean	0.000	0.002 (0.031)
	SD	1.000	1.002 (0.013)
	SK	0.000	0.007 (0.040)
	KR	-1.000	-1.001 (0.041)
	Mean	0.000	0.000 (0.032)
	SD	1.000	1.003 (0.013)
	SK	0.000	0.004 (0.042)
	KR	-1.000	-0.999 (0.036)
	Mean	0.000	0.000 (0.003)
	SD	1.000	1.001 (0.012)
	SK	0.000	0.009 (0.041)
	KR	-1.000	-0.989 (0.025)

Table 2. Average intercorrelations between simulated ability parameters for each ability distribution based on 10 replications and 1000 examinees.

Ability Distribution	Correlation	Expected	Observed Mean (SD)
Normal	r(θ1,θ2)	0.000	0.005 (0.027)
	r(θ1,θ3)	0.000	-0.007 (0.026)
	r(θ2,θ3)	0.000	0.019 (0.033)
Positively Skewed	r(θ1,θ2)	0.000	-0.013 (0.032)
	r(θ1,θ3)	0.000	0.002 (0.025)
	r(θ2,θ3)	0.000	0.020 (0.038)
Platykurtic	r(θ1,θ2)	0.000	0.021 (0.044)
	r(θ1,θ3)	0.000	-0.008 (0.029)
	r(θ2,θ3)	0.000	-0.006 (0.020)
Normal	r(θ1,θ2)	0.600	0.591 (0.023)
	r(θ1,θ3)	0.600	0.603 (0.018)
	r(θ2,θ3)	0.600	0.593 (0.022)
Positively Skewed	r(θ1,θ2)	0.600	0.577 (0.022)
	r(θ1,θ3)	0.600	0.575 (0.023)
	r(θ2,θ3)	0.600	0.573 (0.023)
Platykurtic	r(θ1,θ2)	0.600	0.588 (0.019)
	r(θ1,θ3)	0.600	0.593 (0.015)
	r(θ2,θ3)	0.600	0.582 (0.013)

Table 3. Decsription of simulated data sets based on 10 replications and 1000 simulated examinees for 40 items .

Ability Distribution	Raw Score		Kuder-Richardson Reliability	Traditional Difficulty		Item-Total Test Biserial	
	Mean	SD		p	SD	Mean	SD
Normal (1-dimensional)	22.110	8.241	0.923	0.553	0.240	0.594	0.141
Positively Skewed (1-dimensional)	21.927	8.415	0.925	0.548	0.237	0.592	0.144
Platykurtic (1-dimensional)	22.045	8.682	0.929	0.551	0.231	0.615	0.151
Normal (3-dimensional; corr=0.00)	22.216	8.066	0.934	0.555	0.292	0.671	0.149
Positively Skewed (3-dimensional; corr=0.00)	21.893	8.051	0.934	0.547	0.293	0.656	0.139
Platykurtic (3-dimensional; corr=0.00)	22.092	8.181	0.935	0.552	0.290	0.669	0.153
Normal (3-dimensional; corr=0.60)	22.327	10.386	0.956	0.558	0.246	0.787	0.137
Positively Skewed (3-dimensional; corr=0.60)	21.678	10.385	0.957	0.542	0.247	0.770	0.138
Platykurtic (3-dimensional; corr=0.60)	22.230	10.925	0.959	0.556	0.233	0.804	0.137

Table 4. Average root mean squared error (RMSE) between estimates of item discrimination (a), difficulty (b), and guesing (c) and their true values, and averages (a_{avg}, b_{avg},) based on 10 replications and 1000 simulated examinees for 40 items.

Ability Distribution	RMSE ($\hat{a}$, a_1)	RMSE ($\hat{a}$, a_2)	RMSE ($\hat{a}$, a_3)	RMSE ($\hat{a}$, a_{avg})	RMSE ($\hat{\theta}$, b_1)	RMSE ($\hat{\theta}$, b_2)	RMSE ($\hat{\theta}$, b3)	RMSE ($\hat{\theta}$, b_{avg})	RMSE ($\hat{c}$, c)
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Normal (1-dimensional)	0.164 (0.016)				0.149 (0.081)				0.044 (0.005)
Positively Skewed (1-dimensional)	0.208 (0.017)				0.213 (0.037)				0.073 (0.005)
Platykurtic (1-dimensional)	0.167 (0.100)				0.154 (0.068)				0.051 (0.007)
Normal (3-dimensional; corr=0.00)	0.559 (0.0.63)	0.760 (0.048)	0.588 (0.048)	0.563 (0.063)	0.823 (0.021)	0.921 (0.027)	0.778 (0.032)	0.558 (0.055)	0.036 (0.006)
Positively Skewed (3-dimensional; corr=0.00)	0.507 (0.038)	0.686 (0.042)	0.541 (0.041)	0.488 (0.045)	0.856 (0.028)	0.959 (0.091)	0.801 90.020)	0.620 (0.029)	0.039 (0.014)
Platykurtic (3-dimensional; corr=0.00)	0.512 (0.035)	0.719 (0.043)	0.566 (0.045)	0.520 (0.042)	0.816 (0.033)	0.912 (0.062)	0.771 (0.019)	0.574 (0.038)	0.037 (0.008)
Normal (3-dimensional; corr=0.60)	1.181 (0.079)	1.327 (0.069)	1.205 (0.076)	1.232 (0.075)	0.681 (0.011)	0.792 (0.012)	0.625 (0.028)	0.305 (0.027)	0.027 (0.005)
Positively Skewed (3-dimensional; corr=0.60)	1.132 (0.047)	1.274 (0.054)	1.159 (0.047)	1.180 (0.008)	0.687 (0.052)	0.809 (0.011)	0.632 (0.011)	0.297 (0.059)	0.034 (0.002)
Platykurtic (3-dimensional; corr=0.60)	1.224 (0.057)	1.363(0.091)	1.254 (0.053)	1.273 (0.053)	0.671 (0.010)	0.782 (0.023)	0.613 (0.009)	0.263 (0.0320	0.027 (0.007)

Table 5. Average root mean squared error (RMSE) term between ability estimate and true value(θ), and average (θ_{avg}) based on 10 replications and 1000 simulated examinees for 40 items.

Ability Distribution	RMSE $(\hat{\theta}, \theta_1)$ Mean (SD)	RMSE $(\hat{\theta}, \theta_2)$ Mean (SD)	RMSE $(\hat{\theta}, \theta_3)$ Mean (SD)	RMSE $(\hat{\theta}, \theta_{avg})$ Mean (SD)
Normal (1-dimensional)	0.268 (0.009)			
Positively Skewed (1-dimensional)	0.287 (0.005)			
Platykurtic (1-dimensional)	0.255 (0.007)			
Normal (3-dimensional; corr=0.00)	0.869 (0.019)	0.948 (0.033)	0.922 (0.013)	0.448 (0.011)
Positively Skewed (3-dimensional; corr=0.00)	0.808 (0.021)	0.953 (0.026)	0.910 (0.019)	0.455 (0.011)
Platykurtic (3-dimensional; corr=0.00)	0.825 (0.014)	0.922 (0.025)	0.940 (0.026)	0.436 (0.010)
Normal (3-dimensional; corr=0.60)	0.550 (0.022)	0.593 (0.032)	0.563 (0.009)	0.261 (0.020)
Positively Skewed (3-dimensional; corr=0.60)	0.563 (0.006)	0.594 (0.006)	0.576 (0.011)	0.261 (0.018)
Platykurtic (3-dimensional; corr=0.60)	0.534 (0.120)	0.587 (0.016)	0.566 (0.015)	0.233 (0.016)

Table 6. Results of 3-way ANOVA for testing significance of test dimensionality (1-, 3-dimension), ability distribution(normal, positively skewed, platykurtic), and their interaction when underlying abilities are (un)correlated (0, 0.6), if root mean squared error (RMSE) is used as a dependent variable.

Source of Variation	RMSE			
	average item discrimination a_{avg}	average item difficulty b_{avg}	item guessing c	average ability θ_{avg}
	F-value (sign.)	F-value (sign.)	F-value (sign.)	F-value (sign.)
correlation=0.0				
Test Dimensionality	265.30 (<.001)	111.76 (<.001)	13.06 (<.001)	5465.00 (<.001)
Ability Distribution	0.33 (.720)	0.70 (.496)	3.49 (.032)*	38.10 (<.001)*[a,b,c]
Interaction	2.69 (.070)	0.05 (.949)	2.49 (.085)	3.30 (.037)
correlation=0.6				
Test Dimensionality	1508.82 (<.001)	27.23 (<.001)	30.33 (<.001)	106.72 (<.001)*
Ability Distribution	0.35 (.705)	1.50 (.226)	5.57 (.004)*	105.43 (<.001)*[a,b,c]
Interaction	2.20 (.113)	0.88 (.414)	1.88 (.154)	11.80 (<.001)

a: Normal ability distribution is significantly different from positively skewed ability distribution at 0.05 according to Scheffe procedure
b: Normal ability distribution is significantly different from playtkurtic ability distribution at 0.05 according to Scheffe procedure
c:Positively skewed ability distribution is significantly different from platykurtic ability distribution at 0.05 according to Scheffe procedure

Table 7. Results of 3-way ANOVA for testing significance of intercorrelations between ability dimension (0, 0.6) , ability distribution (normal, positively skewed, platykurtic), and their interaction when test dimension is 3, If root mean squared error (RMSE) is used as a dependent variable.

Source of Variation	RMSE			
	Item discrimination a_{avg}	item difficulty b_{avg}	item guessing c	ability θ_{avg}
	F-value (sign.)	F-value (sign.)	F-value (sign.)	F-value (sign.)
Correlation	487.10 (<.001)	51.65 (<.001)	4.16 (.042)	6696.06 (<.001)
Ability Distribution	1.74 (.178)	0.32 (.726)	0.72 (.488)	37.51 (<.001)[b,c]
Interaction	0.61 (.544)	0.08 (.925)	0.21 (.814)	3.57 (.028)

a: Normal ability distribution is significantly different from positively skewed ability distribution at 0.05 according to Scheffe procedure
b: Normal ability distribution is significantly different from playtkurtic ability distribution at 0.05 according to Scheffe procedure
c:Positively skewed ability distribution is significantly different from platykurtic ability distribution at 0.05 according to Scheffe procedure

Figure 1. Interaction effect between test dimension and ability distribution if underlying ability dimensions are uncorrelated and correlated.

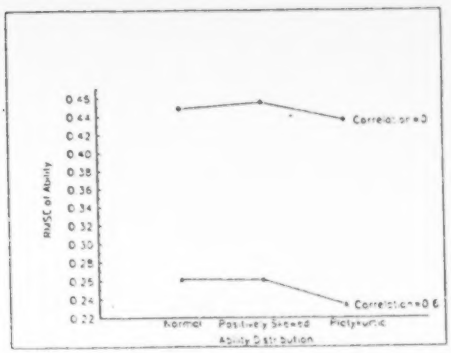
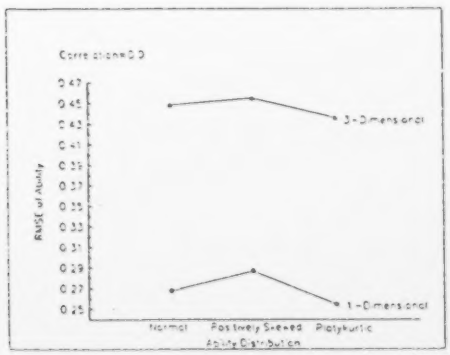
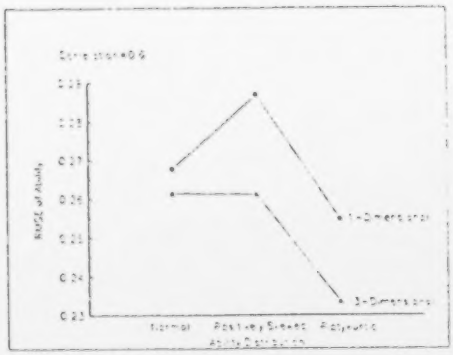


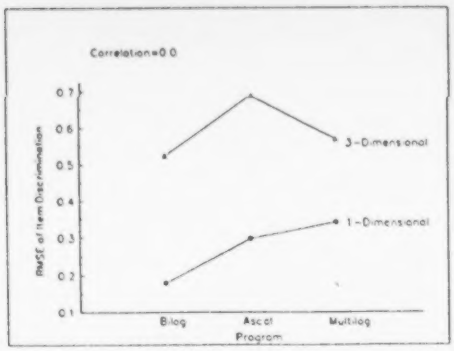
Figure 2. Interaction effect between degree of intercorrelation among ability diemnsions and ability distribution if a test is 3-dimensional.

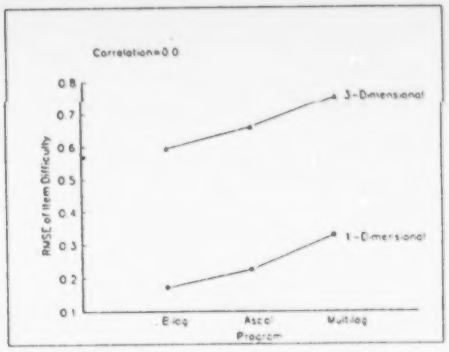


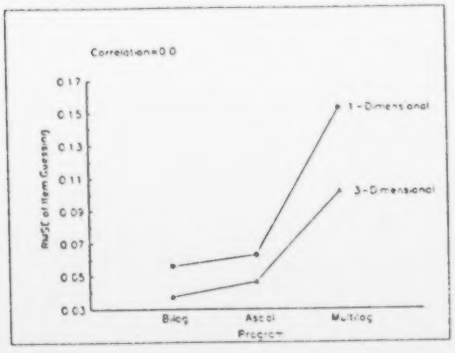


BEST COPY AVAILABLE

Figure 3. Comparison of BILOG, ASCAL, and MULTILOG for 1- and 3-dimensional tests with respect to RMSE of item parameters and ability if the degree of intercorrelation among abilities are 0.







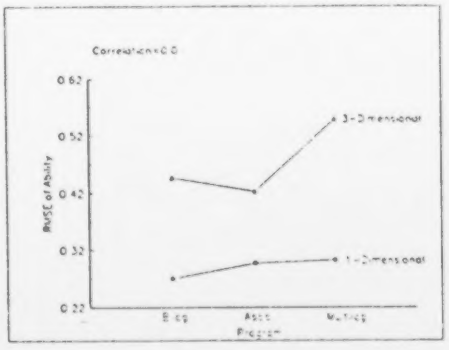


Figure 4. Comparison of BILOG, ASCAL, and MULTILOG for 1- and 3-dimensional tests with respect to RMSE of item parameters and ability if the degree of intercorrelation among abilities are 0.6.

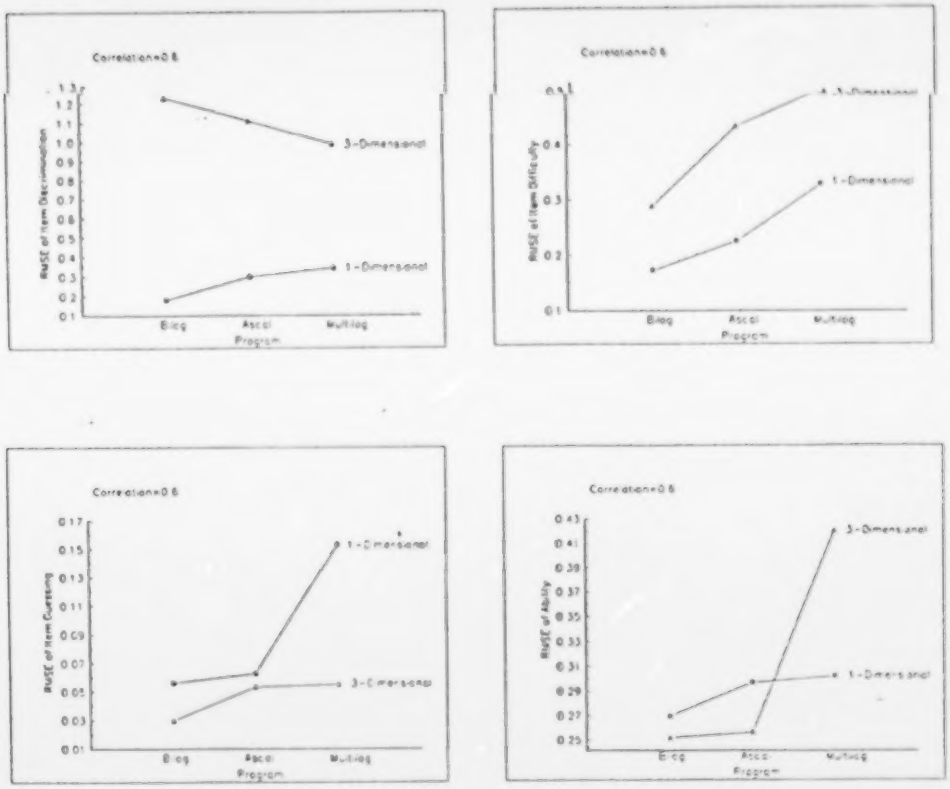
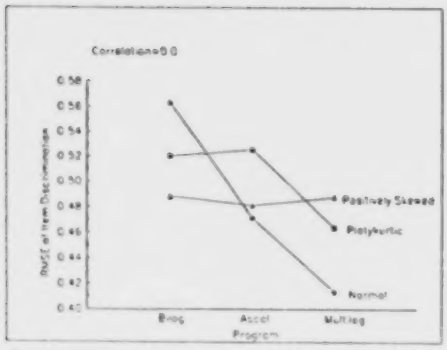
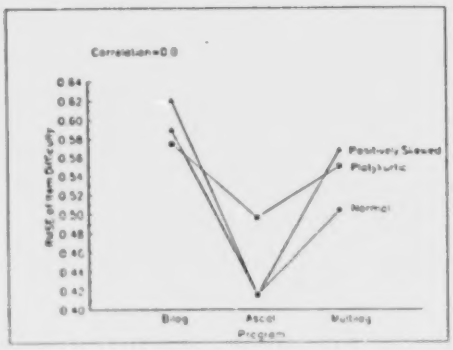
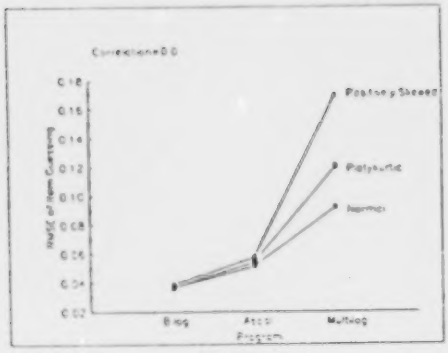


Figure 5. Comparison of BILOG, ASCAL, and MULTILOG for ability distribution with respect to RMSE of item parameters and ability if the degree of intercorelation among abilities is 0.







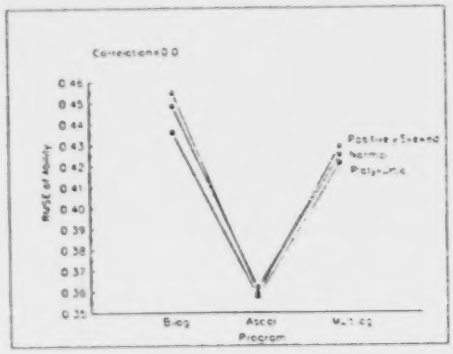
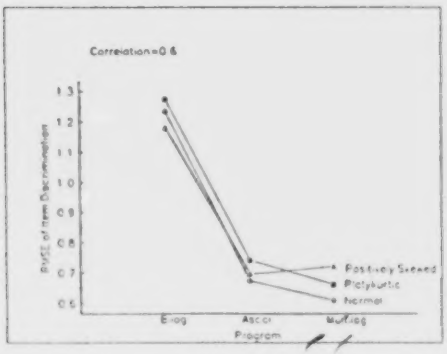
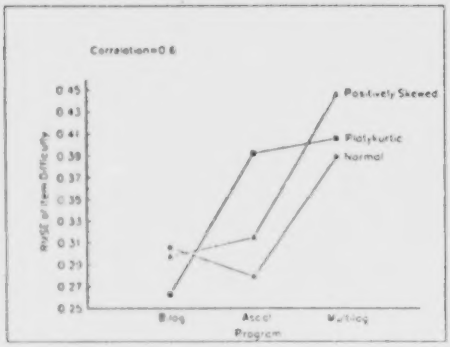
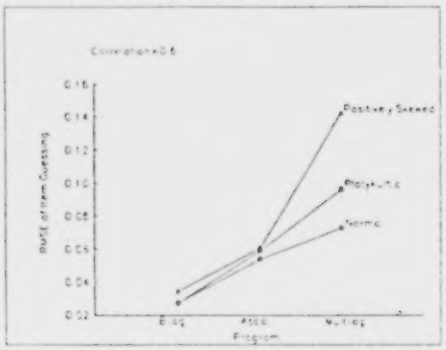


Figure 6. Comparison of BILOG, ASCAL, and MULTILOG for ability distribution with respect to RMSE of item parameters and ability if the degree of intercorelation among abilities is 0.6.







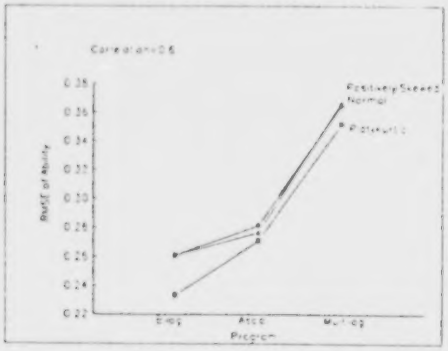


Figure 7. Comparison of BILOG, ASCAL, and MULTILOG for degree of intercorrelation among ability dimensions with respect to RMSE of item parameters and ability.